Les entités nommées : des clés linguistiques pour la conceptualisation

Nouha Omrane[1], Adeline Nazarenko[2], Sylvie Szulman[3]

LIPN (Laboratoire d'Informatique de Paris Nord)
Université Paris 13 & CNRS (UMR 7030)
nouha.omrane@lipn.univ-paris13.fr
adeline.nazarenko@lipn.univ-paris13.fr
sylvie.szulman@lipn.univ-paris13.fr

Résumé : Partir de textes pour construire des ontologies présente de nombreux avantages. Cela permet notamment de produire des ontologies enrichies d'informations lexicales qui sont précieuses pour toutes les applications d'accès au contenu. La construction d'ontologies à partir de textes est un domaine qui ne cesse d'évoluer. Même si le processus de construction d'ontologies à partir de textes n'est pas entièrement automatique, l'ingénieur de la connaissance peut être guidé durant le processus de construction. Dans cet article, nous montrons que la détection des entités nommées peut servir à enrichir une ontologie existante ou à démarrer une conceptualisation et pas seulement à peupler une ontologie. Ce propos est illustré par deux cas d'usage portant sur des documents réglementaires et nous évaluons notre approche en comparant les ontologies construites par rapport à des références.
Mots clés : construction d'ontologies, entité nommée, conceptualisation.

1. Introduction

La conception d'ontologies nécessite à la fois des ressources et un important travail de modélisation. Différentes sources peuvent être utilisées. L'ontologie peut être construite à partir de bases de données, de textes, d'ontologies préexistantes, d'interviews d'experts ou même à partir de diagrammes et de tableaux. Une fois ces sources identifiées, l'ingénieur de la connaissance doit repérer quels sont les éléments les plus pertinents à prendre en compte, les définir, analyser les relations existant entre eux, et les représenter de manière formelle.

Parmi les ressources disponibles, nous nous intéressons aux textes. Les documents ont le mérite d'être souvent faciles d'accès et ils reflètent une connaissance générale du domaine même si les auteurs ne l'explicitent pas toujours. A l'inverse, les experts du domaine sont peu disponibles pour des longues interviews et ils n'ont souvent qu'une vue partielle du domaine à modéliser. Un autre avantage de l'utilisation du texte comme source de connaissance est lié à l'exploitation ultérieure de l'ontologie. Dans le cas où

l'ontologie sert à annoter les documents pour des applications de recherche d'information ou pour indexer des documents, il est utile de garder le lien avec le texte. Les informations textuelles recueillies au cours du processus de conceptualisation servent à lier l'ontologie aux documents.

Dans le cas où la source de la connaissance est le texte, l'ingénierie de la connaissance fait appel à des outils de traitement automatique de la langue (TAL) pour analyser le texte. Ils permettent d'en extraire des éléments d'information qui reflètent le domaine. Dans cet article, nous nous intéressons en particulier aux termes et entités nommées que nous considérons *a priori* comme des éléments pertinents pour la création de l'ontologie. Cela concerne notamment la phase de conceptualisation permet de structurer le modèle conceptuel (T-Box en logique de description). Contrairement aux travaux qui considèrent les entités nommées comme des instances destinées à peupler l'ontologie (A-Box dite aussi base de connaissances) (Maynard *et al.*, 2008), nous proposons de considérer les entités nommées comme des marqueurs textuels à prendre en compte dès la phase de conceptualisation.

Nous montrons coment la méthode TERMINAE décrite dans (Aussenac-Gilles *et al.*, 2008) peut être enrichie par la prise en compte des entités nommées en parallèle des termes et ce, dès le début du processus de conceptualisation [1]. Nous nous appuyons sur des cas d'usage réels visant à élaborer des systèmes d'aide à la décision. Dans ce contexte, on s'appuie sur des textes réglementaires pour élaborer des ontologies de domaine. Cela permet de modéliser le vocabulaire conceptuel à utiliser pour formuler les règles métier sur lesquelles repose la prise de décision [2].

L'article est composé de quatre sections. La première présente l'état de l'art en matière de méthodes de construction d'ontologies à partir de textes et le rôle des entités nommées dans ce processus. La deuxième décrit la méthode TERMINAE enrichie par l'identification des entités nommées. La troisième section présente deux expérimentations faites dans le domaine règlementaire. Nous évaluons notre approche dans la dernière section en lien avec ce contexte applicatif.

2. Méthodes de construction d'ontologies à partir de textes

Les méthodes de construction d'ontologies à partir de textes se différencient par le type d'unités textuelles qu'elles exploitent (mot, terme, entité nommée [3]) et par leur degré d'automatisation.

1. L'outil TERMINAE (Aussenac-Gilles *et al.*, 2008) a été redéveloppé et enrichi (voir http ://www-lipn.univ-paris13.fr/ szulman/logi), mais nous n'abordons ici que les aspects méthodologiques.

2. Les documents utilisés sont extraits des cas d'usage étudiés dans le cadre du projet ONTORULE (FP7 231875).

3. Beaucoup de travaux se sont intéressés à l'extraction des relations mais nous n'intéressons pas dans cet article à l'étude des propriétés.

2.1. Les approches distributionnelles basées sur les mots

Dans la famille des approches distributionnelles, on peut distinguer les méthodes présentées comme automatiques des méthodes semi-automatiques.

L'approche distributionnelle, de manière générale, est héritée de (Harris *et al.*, 1989) et vise à extraire des concepts à partir de l'analyse des mots et de leur distribution dans le texte. Elle repose sur l'hypothèse que les mots similaires apparaissent dans des contextes similaires et elle a été proposée au départ sur des corpus de domaines spécialisés sur lesquels elle est supposée être plus fiable. Les classes sémantiques sont créées à partir des groupements de mots ayant des distributions proches. L'ontologie est construite à partir de ces classes sémantiques qui forment des concepts candidats.

Le plugin de l'éditeur d'ontologie Protégé OntoLp propose à l'ingénieur de la connaissance des concepts candidats ainsi que des relations à partir des indices textuels. L'utilisateur n'est sollicité que pour filtrer les mots et les groupes de mots formés comme rapporté dans l'expérimentation de (Lopes & Vieira, 2009). Text2Onto (Cimiano & Völker, 2005), qui se base sur la plate-forme GATE, exploite des outils de TAL pour la lemmatisation, l'extraction de termes et l'apprentissage de concepts et relations sémantiques. L'ensemble de ce éléments permet de créer une ébauche d'ontologie. Néanmoins l'ontologie construite nécessite un travail de correction par l'ingénieur de la connaissance comme précisé dans (Wang *et al.*, 2006).

D'autres travaux conçoivent la construction d'ontologies de manière incrémentale et coopérative. La même approche distributionnelle est mise en œuvre mais d'une manière différente : la construction d'ontologies est réalisée par un ou plusieurs intervenants. Les systèmes ASIUM (Faure & Nédellec, 1999) et OntoGen [4] sont des outils de groupement supervisé respectivement ascendant et descendant. ASIUM considère les groupes formés par les mots comme des concepts candidats et procède de manière ascendante allant des concepts les plus spécifiques vers les concepts les plus génériques. C'est à l'ingénieur de la connaissance de valider les groupes formés à chaque étape du processus de groupement. A l'opposé d'ASIUM, l'outil OntoGen propose des classes (groupes) de concepts généraux au domaine puis d'une manière descendante spécifie d'autres classes plus précises. L'objectif, dans ce cas, est de construire des classes permettant d'indexer des documents, autrement dit de construire un plan de classification.

L'inconvénient majeur de la première approche est que l'ingénieur de la connaissance valide manuellement le résultat obtenu à la fin du processus de construction de l'ontologie sans pouvoir intervenir durant le processus de construction.

2.2. L'approche terminologique

Une terminologie (Meyer *et al.*, 1992; Aussenac-Gilles *et al.*, 1995) est souvent définie comme l'ensemble des mots ou termes relatifs à un sujet particulier ou à un domaine

4. http ://ontogen.ijs.si/

d'activité, dans l'hypothèse où un domaine de connaissance est caractérisé par un langage propre dont les termes constituent le vocabulaire.

(Lerat, 2009) définit la notion de terme comme : « le nom donné dans une langue à une entité conceptualisée par une communauté de travail. Cette dénomination est souvent un nom ou un groupe nominal, mais elle peut aussi appartenir à une nomenclature alphanumérique, une unité définie dans les textes de spécialité ». Les termes ont généralement un sens précis dans le domaine auquel ils appartiennent, ils sont moins ambigus que les mots courants et présentent moins de variation de forme. Pourtant, il n'existe pas de critère formel permettant de déterminer si un mot ou un syntagme a ou non une valeur terminologique et les outils de TAL ne peuvent pas détecter de manière fiable tous les termes pertinents d'un domaine. Pour avoir un résultat de qualité, il faut faire intervenir un expert pour valider les termes importants du domaine considéré.

Dans le processus de construction d'ontologies, les termes servent souvent à identifier les éléments pertinents du domaine (Cimiano, 2006) et à démarrer la construction de l'ontologie même si les concepts ne sont pas pour autant créés automatiquement à partir des termes. Dans la méthode TERMINAE, la construction des concepts à partir des termes est assurée au départ par un processus de filtrage et de sélection au sein duquel des choix de modélisation sont faits par l'ingénieur de la connaissance.

2.3. Le rôle des entités nommées dans la construction d'ontologie

Un autre ensemble de travaux vise à peupler les ontologies une fois qu'elles sont construites.

Le peuplement d'une ontologie à partir de textes est un champ d'application de l'extraction d'information (Buitelaar *et al.*, 2005), si l'on considère que la structure d'information (ou formulaire) à remplir est en réalité une structure ontologique qui sert de grille d'interprétation dans le processus d'extraction (Bontcheva & Cunningham, 2003; Nédellec *et al.*, 2009). Une tâche très classique d'extraction d'information consiste à repérer des unités textuelles particulières qu'on appelle « entités nommées ». Il s'agit de reconnaître les noms propres (personnes, lieux, organisations) mais aussi les expressions temporelles (dates, durées, horaires) et les noms de quantités (monétaires, unités de mesure, pourcentages) puis de les étiqueter en leur associant un type sémantique. La complexité de la tâche est liée en grande partie à la richesse de la typologie des entités à retrouver. De quelques grands types généraux (lieux, personnes, dates), on est passé à des typologies de 200 types (Sekine & Nobata, 2004), le succès de cette tâche (Nadeau & Sekine, 2007) s'expliquant notamment par la diversité de ses champs d'application (systèmes de question-réponse, d'indexation, d'intégration de données issues des textes ou d'anonymisation).

L'utilisation de ces outils d'extraction d'entités nommées pour le peuplement des ontologies (Magnini *et al.*, 2006) repose sur l'hypothèse que la mention d'un nom d'entité de type T dans un texte révèle l'existence de cette entité et permet de créer une instance du concept T dans l'ontologie. (Tanev & Magnini, 2008) cherchent ainsi à col-

lecter des instances de lieux et de personnes à partir de Wikipedia en les catégorisant précisément [5]. Des problèmes d'ambiguïté dans la détermination du type T et de rapprochement des noms correspondant à la même entité peuvent se poser mais le principe de base reste simple.

Le fait que les entités nommées soient largement utilisées pour le peuplement d'ontologies ne signifie pas pour autant qu'elles soient à négliger pour la conceptualisation et le fait que la conceptualisation repose traditionnellement et prioritairement sur l'analyse terminologique n'interdit pas d'exploiter d'autres indices textuels.

Le cloisonnement des processus de conceptualisation et de peuplement paraît en réalité assez arbitraire, car les choix de conceptualisation au niveau de la T-Box se font souvent en anticipant le nombre et la nature des instances qui devront figurer dans la A-Box et sur lesquelles le raisonnement ontologique est destiné à reposer. La question se pose donc de rapprocher les deux processus et d'exploiter les techniques de peuplement dès la phase de conceptualisation.

3. Une méthode TERMINAE enrichie

Aucune des méthodes citées précédemment ne permet de construire automatiquement une ontologie à partir de textes. L'intervention humaine est nécessaire : l'ingénieur de la connaissance extrait les connaissances pertinentes des textes et construit un modèle du domaine adapté à l'application visée. La méthode TERMINAE permet à l'ingénieur de la connaissance de s'appuyer sur le matériau linguistique pour construire un tel modèle.

3.1. Présentation de la méthode

La méthode TERMINAE est une méthode de construction d'ontologies à partir de textes qui décompose le processus d'acquisition en trois niveaux terminologique, terminoontologique et conceptuel (figure 1 [6]), un corpus d'acquisition servant de socle à l'ensemble [7]. La transition du texte à l'ontologie n'est pas automatique, c'est pourquoi l'acquisition est un processus interactif dans la méthode TERMINAE.

3.1.1. L'analyse terminologique

La première étape (notée 1 dans la figure 1) est automatique et permet d'extraire du texte des unités textuelles qui semblent fonctionner comme des termes pour le domaine considéré. Les résultats étant cependant bruités, l'ingénieur de la connaissance doit sélectionner les unités qui lui paraissent les plus pertinentes. L'outil TERMINAE intègre

5. Cinq sous-types de lieux et de personnes sont pris en considération : MOUNTAIN, LAKE, RIVER, CITY, COUNTRY ; STATESMAN, WRITER, ATHLETE, ACTOR, INVENTOR.

6. Dans cet article, on ne s'intéresse qu'à la partie supérieure du schéma.

7. La première étape consiste à créer un corpus représentatif du domaine. Ce point n'est pas traité ici.

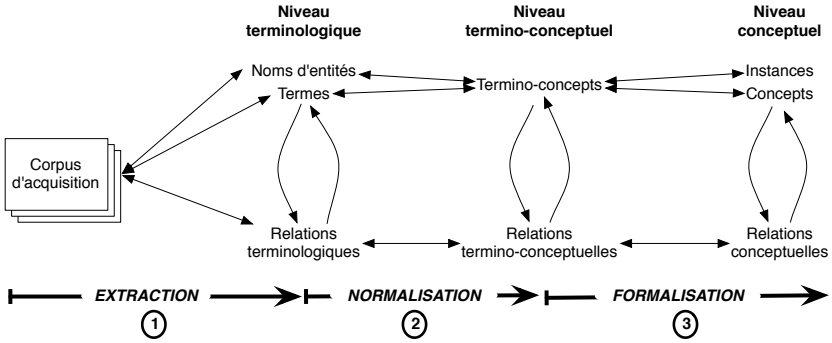

FIGURE 1: Les trois niveaux de connaissance et les trois étapes de la méthodologie de TERMINAE

désormais des resultats d'outils de reconnaissance d'entités nommées [8] en plus de l'extraction de termes [9]. Les résultats de l'analyse linguistique se visualisent sous la forme d'une liste de termes candidats et d'entités nommées. L'ingénieur de la connaissance doit filtrer la liste obtenue et regrouper certaines unités. Les unités lexicales extraites peuvent être mal formées car les documents du domaine contiennent des particularités orthographiques, syntaxiques et lexicales que les outils de TAL génériques ne peuvent prendre en considération. Certaines unités lexicales ne sont pas pertinentes pour le domaine. A l'issue de cette phase, la liste obtenue contient les termes et entités nommées validés et considérés comme pertinents par l'ingénieur de la connaissance.

Parmi les unités retenues, l'ingénieur de la connaissance peut ensuite identifier des unités synonymes (qui décrivent un même sens ou renvoient à une même entité du monde) et les regrouper sous une forme canonique. Il peut s'agir de variante de natures diverses : graphique (*NY* vs. *New York*), morphologique (*member* vs. *members*) ou bien morpho-syntaxique (*program member* vs. *member of the program*). Durant l'étape terminologique, l'analyse de la liste obtenue n'est généralement pas exhaustive.

3.1.2. La création des termino-concepts

Nous définissons le termino-concept comme un terme désambiguïsé dont le sens est défini par son usage dans le corpus. La liste des unités lexicales pertinentes est transformée en un réseau de termino-concepts structuré par des liens hiérarchiques (des relations de généricité-spécificité). Durant cette étape (notée 2 dans la figure 1), l'ingénieur de la connaissance commence à normaliser le réseau sémantique en vue de la création d'un modèle du domaine plus proche du conceptuel que du linguistique (modèle semi-formel).

Ce processus de normalisation sémantique permet à nouveau de regouper les termes

8. ANNIE est une chaîne de traitement de la plateforme Gate, http ://gate.ac.uk/

9. `http://search.cpan.org/~thhamon/Lingua-YaTeA-0.5/`

synonymes en les associant à un unique termino-concept et de désambiguïser les termes ambigus, car l'ingénieur de la connaissance crée un termino-concept pour chaque sens pertinent de terme.

Le résultat de la phase termino-conceptuelle est un réseau terminologique normalisé qui décrit le domaine considéré et s'apparente à un thésaurus. Le modèle obtenu peut être exporté à partir de l'outil TERMINAE sous la forme d'un fichier SKOS [10], le standard W3C utilisé pour representer ce type de structure de connaissance. A chaque termino-concept est attribué un ou plusieurs termes, des synonymes, des définitions, des notes et des relations de généricité-spécificité et d'associativité.

3.1.3. La création des concepts

Durant cette phase (notée 3 dans la figure 1) de la méthode TERMINAE, le réseau obtenu à l'étape précédente est formalisé en ontologie. Là non plus, la transformation d'un termino-concept en un concept décrit en OWL n'est pas automatique. L'ingénieur de la connaissance est amené à faire des choix de formalisation. Plusieurs questions importantes doivent être traitées :

- la distinction entre concepts généraux et concepts individuels (classes et instances) n'est pas établie au niveau termino-conceptuel ;
- il faut organiser les concepts dans une structure hiérarchique, pour assurer les traitements ultérieurs comme les mécanismes d'inférence et de subsomption ;
- l'ontologie de domaine doit souvent être rattachée à une ontologie générique ou à une ou plusieurs ontologie(s) existantes. Il s'agit de réutiliser l'information existante afin d'économiser l'effort et le temps passés pour construire la couche supérieure de l'ontologie.

L'éditeur d'ontologies qui permet ce travail est celui du plugin NEON TOOLKIT ONTOLOGY (version 2.4) qui est intégré dans l'outil TERMINAE.

3.2. La conceptualisation des entités nommées

Prendre en considération les entités nommées dès le début du processus de construction de l'ontologie soulève la question de leur rôle dans la phase de conceptualisation. C'est à l'ingénieur de la connaissance de faire des choix de modélisation qui sont les plus appropriés aux contraintes du domaine. Contrairement aux approches qui tendent à peupler les ontologies en dérivant automatiquement un type ontologique (instance) à partir de sa catégorie linguistique (entité nommée), TERMINAE propose de relier n'importe quel type ontologique à n'importe quel élément linguistique à l'aide de termino-concepts indistincts. L'ingénieur de la connaissance crée des termino-concepts pour tous les termes et entités nommées pertinents et désambiguïsés du domaine. Puis il associe à ces derniers des concepts, instances ou relations conceptuelles au niveau concep-

10. SKOS (Simple Knowledge Organisation System) est un standard pour la représentation de thésaurus ou de taxonomies pour le web sémantique. http ://www.w3.org/2004/02/skos/

tuel. Nous ne faisons pas de distinction entre la T-Box et la A-Box au niveau termino-conceptuel. Les choix d'un type ontologique se fait au moment de la transition des termino-concepts vers les éléments ontologiques.

Dans certains contextes spécifiques, les termes font référence à une entité du monde (*mon université*, *le président de l'état*) et les entités nommées peuvent être modélisées comme des concepts prenant eux-mêmes des instances. Par exemple, suivant l'application visée, un gène peut être modélisé comme une instance d'un concept **Gène** ou bien comme un concept qui représente un type spécifique d'instances de gènes.

Le travail précis d'analyse des termes ou entités nommées puis des termino-concepts se fait dans TERMINAE à l'aide de fiches terminologiques ou termino-conceptuelles qui regroupent dans une même vue toutes les informations obtenues automatiquement ou manuellement pour une unité linguistique (terme ou entité nommée) ou termino-conceptuelle. La fiche terminologique (figure 2) décrit toutes les caractéristiques d'un terme ou d'une entité nommée et l'associe à un ou plusieurs termino-concepts suivant que l'unité linguistique décrit un ou plusieurs sens pertinents à conceptualiser dans la phase suivante. Pour une unité linguistique sélectionnée dans la liste affichée à gauche.

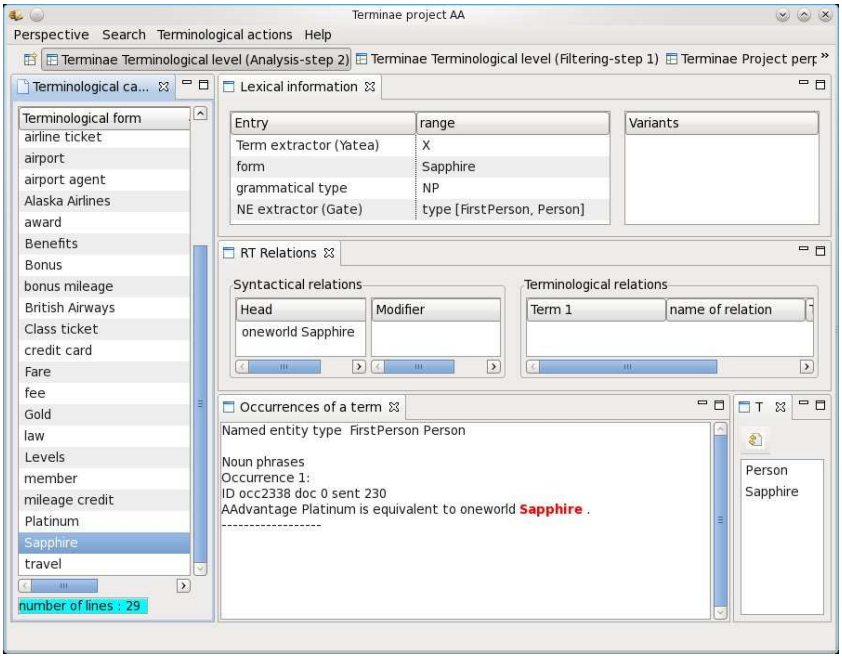

FIGURE 2: La fiche terminologique de l'entité nommée *Sapphire*

La fiche terminologique de droite indique si l'élément sélectionné est un terme, une entité nommée ou bien les deux (onglet « Lexical information »). L'ingénieur de la connaissance intervient à ce stade pour valider la fiche créée : en parcourant la liste des occurrences de l'unité linguistique en question (onglet « Occurrences of a term »),

il peut vérifier si elle possède un ou plusieurs sens, choisir le(s) pertinent(s) voire en ajouter. Dans l'exemple de la figure 2, la fiche terminologique décrit une entité nommée *Sapphire* qui n'a qu'une occurrence dans le texte. Nous présentons plus loin le cas d'usage correspondant mais il s'agit d'une catégorie de voyageurs qui est importante à prendre en considération, si bien qu'il est nécessaire de créer un termino-concept ***Sapphire***. Les entités nommées sont des unités linguistiques pertinentes qui possèdent des types sémantiques spécifiques attribués par les outils de reconnaissance d'entités nommées. En l'occurrence, le type sémantique PERSON a été associé à l'entité nommée *Sapphire*. Comme ce type sémantique peut lui même être pertinent pour le domaine à modéliser, cette information est sauvegardée dans la fiche terminologique de l'entité correspondante sous la forme d'un termino-concept. L'entité nommée *Sapphire* est ainsi reliée à deux termino-concepts : ***Sapphire*** et ***Person***.

La figure 3 présente la fiche termino-conceptuelle décrivant le termino-concept *Sapphire* créé à partir de l'entité nommée précédente qui possède un synonyme (*AAdvantage platinum*, onglet « Synonyms »). L'ingénieur de la connaissance saisit une définition du terme ou de l'entité nommée en langue naturelle (onglet « Definition »).

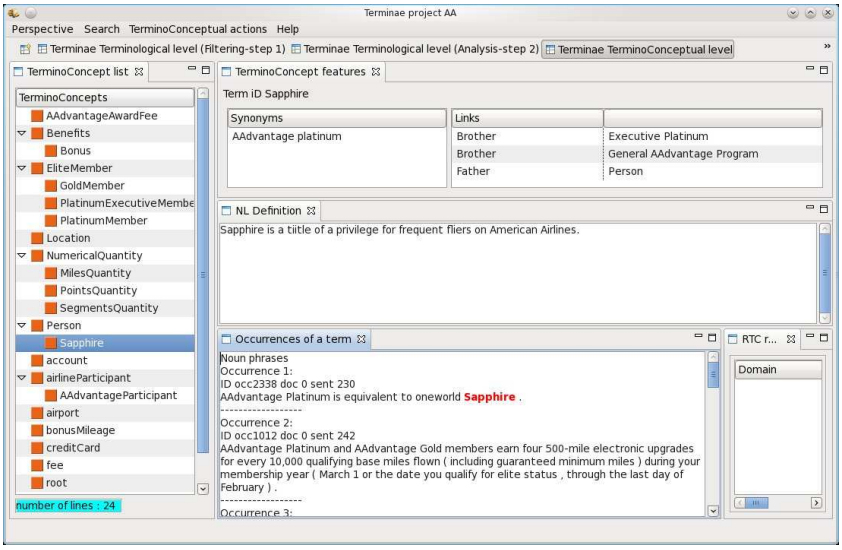

FIGURE 3: La fiche termino-conceptuelle du termino-concept ***Sapphire***

3.3. Démarrer la phase de conceptualisation

Les méthodologies d'acquisition ne proposent pas de stratégie pour l'exploration des données. Les grandes masses de données (longues listes de termes et d'entités nommées) sont difficiles à gérer, sauf par des approches distributionnelles et il est difficile pour l'ingénieur de la connaissance de savoir comment démarrer. La démarche consistant à partir d'une analyse fine et à la compléter au niveau terminologique pour ensuite passer

au niveau termino-conceptuel puis au niveau ontologique n'est pas réaliste. L'analyse terminologique est d'autant plus fructueuse qu'on peut visualiser rapidement même un fragment de l'ontologie en cours de construction. Suivant le corpus et le cas d'usage traité, plusieurs stratégies peuvent être envisagées :

- L'utilisation d'une ressource externe (ontologie existante, thésaurus, etc.) aide à mettre l'accent sur des éléments linguistiques ou conceptuels qui peuvent être intéressants pour le domaine. L'ingénieur de la connaissance se sert de ces éléments pour explorer le corpus du domaine et repérer d'autres éléments pertinents à conceptualiser. Ainsi la ressource externe permet d'amorcer la conceptualisation.
- Dans certains cas, on peut directement partir d'un modèle conceptuel (une ontologie existante) pour l'enrichir et l'ancrer au texte.
- Dans le cas où il n'existe pas de ressources externes, l'ingénieur de la connaissance se sert du corpus pour trouver les unités lexicales pertinentes pour le domaine. Ceci peut se faire soit en se concentrant sur des parties du texte qui sont importantes à explorer (corpus structuré), soit à partir d'un ensemble de termes considérés comme centraux sur la base de critères de saillance liés par exemple à leur fréquence dans le corpus ou dans des corpus comparables.
- Une autre approche part de l'analyse de faits exprimés dans un corpus. En effet, la détection des entités nommées dans le texte permet d'identifier des éléments factuels. Ces entités permettent de créer les concepts associés et d'explorer les termes et relations figurant dans leur voisinage [11].

La section 4. illustre ces différentes approches dans le domaine réglementaire.

4. Cas d'usage

Nous considérons deux cas d'usage du projet ONTORULE pour lesquels il faut construire des systèmes d'aide à la décision et les bases de règles métiers correspondantes. La méthode TERMINAE permet d'enrichir ou de créer des ontologies de domaine à partir de textes réglementaires (en anglais) et ces ontologies représentent le vocabulaire conceptuel à utiliser pour formuler les règles métier au bon niveau d'abstraction.

Les entités nommées ont été exploitées différemment dans les deux cas d'usage. Dans le premier, elles sont venus enrichir une ontologie qui avait été préalablement construite à partir d'une liste de termes. Dans le deuxième cas, les entités nommées ont été utilisées dès le départ, pour amorcer le processus de conceptualisation.

Même si les documents réglementaires ne contiennent pas toujours beaucoup d'entités nommées, nos expérimentations prouvent l'intérêt de ce type d'unités lexicales pour la phase de conceptualisation.

11. En pratique, le voisinage est défini par la phrase.

4.1. Cas d'usage d'AAdvantage

Dans ce cas d'usage, nous avons créé une ontologie de domaine à partir d'un corpus d'American Airlines qui décrit les règles et conditions d'attribution de « miles » pour des voyageurs [12]. L'analyse de ce corpus [13], donne une liste de 973 termes candidats. Après élimination des termes bruités et regroupement des variantes, nous avons obtenu une nouvelle liste de 634 termes qui nous a permis de créer une première ontologie nommée AA_1. Dans un deuxième temps, nous avons pris les entités nommées en considération. 67 entités nommées ont été extraites [14] du corpus d'AAdvantage. Beaucoup se sont révélées pertinentes.

Certaines entités nommées sont conceptualisées comme concepts. L'entité nommée *Central America* qui, au-delà de sa caractéristique référentielle d'un territoire regroupant des villes (Costa Rica, El Salvador), indique un ensemble d'aéroports spécifiques qui jouent un rôle important dans l'attribution des miles, renvoie à une notion qui n'avait pas émergé lors de la première analyse de la liste initiale. Nous avons donc créé le concept **Central America** et nous lui avons associé des instances (les aéroports).

Nous avons aussi ajouté des instances à un concept existant. Le concept **AAdvantage_Airline_Participant** que nous avions déjà créé à partir du terme *AAdvantage participant* a comme instances l'ensemble des compagnies aériennes qui participent au même programme de fidélité (ex. *American Eagle*, *Japan airways*).

Les types sémantiques associés aux entités nommées ont en outre permis de créer des concepts. Le type ORGANIZATION associé aux compagnies aériennes a ainsi donné le concept **Organization**, père du concept **AAdvantage_Airline_Participant**.

La détection des entités nommées contribue enfin à une meilleure compréhension du domaine et à la création de nouveaux concepts. Par exemple, la reconnaissance des entités nommées *Sapphire* et *Ruby*, que nous avions considérées comme du bruit dans la première analyse terminologique du corpus, a permis de détecter des catégories de statuts de voyageurs avec des règles d'attribution de miles et de bonus différentes. L'exploration des contextes de ces deux entités nommées a permis de mieux comprendre la réglementation et d'intégrer les concepts suivants à l'ontologie initiale (figure 4) :

- **Elite_Member** regroupe tous les statuts qu'un membre peut avoir ;
- **Benefit** décrit les avantages dont un membre peut bénéficier suivant son statut ;
- **Numerical_Quantity** correspond aux différents montants de bonus (points, segments ou miles) que peuvent gagner des voyageurs privilégiés lors de leurs voyages ;
- **Account** est relatif au compte d'un membre contenant les bonus accumulés.

La détection des entités nommées a permis d'identifier des concepts importants pour la modélisation du domaine. Par exemple, selon les cas, un membre ayant le statut Ruby

12. Nous remercions American Airlines d'avoir mis ce corpus à notre disposition.

13. Nous avons utilisé l'extracteur de termes YaTeA (http ://search.cpan.org/ thhamon/Lingua-YaTeA/) qui fait lui-même appel à TreeTagger (http ://www.ims.uni-stuttgart.fr/projekte/corplex/TreeTagger/).

14. Grâce à la chaîne de traitement ANNIE (http ://gate.ac.uk/sale/tao/splitch6.html#chap :annie).

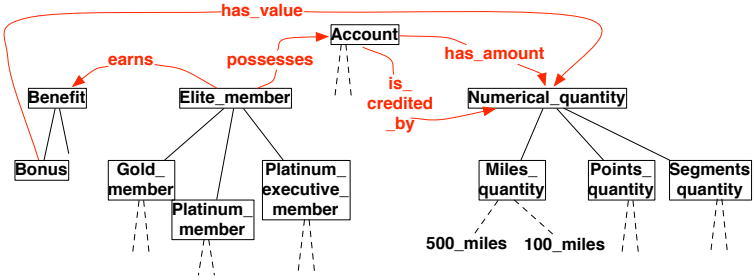

FIGURE 4: Extrait de l'ontologie AA

peut gagner pour un vol aller-retour 25 000 miles, 25 000 points ou 30 segments, ce qui a conduit à introduire des propriétés (**earns**, **possesses**, **has_value**, **has_amount**) avec des restrictions dans l'ontologie.

4.2. Cas d'usage d'Audi

Le corpus Audi est un extrait d'une directive internationale qui décrit les règles et procédures que les véhicules à quatre roues ainsi que leurs équipements doivent satisfaire pour tout ce qui touche aux ceintures de sécurité. Le corpus contient 3 704 mots. La détection de certaines mentions d'entités nommées a permis de démarrer la conceptualisation en contribuant à mettre l'accent sur des termes pertinents du domaine.

Nous avons commencé le processus de conceptualisation en ne considérant au départ que les entités ayant comme type sémantique PERSON, DATE, PERCENT et UNKNOWN.

Nous avions en effet remarqué que ces entités nommées figurent souvent dans des passages réglementaires et qu'elles sont importantes à modéliser parce qu'elle introduisent des valeurs particulières qui jouent des rôles clefs dans les règles métier. Par exemple, certaines dates contraignent des opérations de tests pour la ceinture de sécurité. Grâce à l'exploration des contextes des entités nommées, nous avons identifié les noms des tests (*Calibration test, Dynamic test, Corrosion test*) qui permettent de vérifier si les équipements de voiture respectent les normes ou pas. Les entités nommées de type PERCENT ont permis de créer des propriétés de concepts de domaine qui servent à vérifier si les équipements d'une voiture satisfont les critères de sécurité. Ce type d'entité nommée est mentionné dans le corpus avec des valeurs spécifiques que les méthodes de test doivent vérifier, ce qui a conduit à créer les propriétés correspondantes : **length of a strap** (longueur de la sangle), **temperature**, **humidity** et **salt** (sel).

Les entités nommées de type UNKOWN [15] sont aussi des unités spécifiques au domaine. Ce sont des éléments du vocabulaire de la directive européenne relatif aux procédures appliquées pour la vérification de la ceinture de sécurité. Prenons comme exemple, les entités nommées apparaissant sous la forme de littéraux dans le texte : *M1, N1*, iden-

15. Type est associé aux entités nommées auxquelles aucun type sémantique n'a pu être attribué.

tifient des catégories spécifiques de véhicules. Ces entités nommées sont importantes pour la conceptualisation du domaine car il existe un ensemble de règles et procédures qui peuvent être paramétrées (température, durée, etc.) selon la catégorie du véhicule. De même, les entités nommées telles que *Point A, Point C* font référence à des positions exactes dans la ceinture de sécurité. Ces positions doivent être modélisées puisque les tests et procédures décrits dans la directive vérifient des paramètres qui dépendent de ces positions.

Nous avons ainsi créé 53 concepts de domaine lors de la première itération du processus de conceptualisation en commençant la conceptualisation par l'exploration des seuls contextes des entités nommées pertinentes.

5. Evaluation

Isoler l'apport spécifique des entités nommées est difficile parce qu'elles ne peuvent pas être totalement dissociées des termes dans le travail de conceptualisation mais nous essayons néanmoins de le cerner dans les deux cas d'usage présentés ci-dessus.

Pour évaluer notre approche, nous comparons les résultats que nous obtenons à une « référence » proposée par un expert du domaine. Selon les cas, l'expert a réellement construit une ontologie et nous prenons la liste des concepts comme référence ou bien il a formalisé une liste de règles métier qui nous a permis d'établir un vocabulaire conceptuel, ce qui correspond à une liste de termino-concepts. Pour comparer un résultat à une référence, nous avons utilisé les mesures de précision et rappel qui servent couramment en recherche d'information pour comparer les résultats d'un système avec une référence. Ces mesures se définissent comme suit :

$$\textit{Précision} = \frac{\text{UTP}}{\text{UT}} \quad \textit{Rappel} = \frac{\text{UTP}}{\text{UP}}$$

où UP, UT ou UTP sont respectivement le nombre d'unités pertinentes (*i.e.* figurant dans la référence), le nombre d'unités trouvées (*i.e.* figurant dans le résultat) et le nombre d'unités à la fois trouvées et pertinentes. Selon la nature de la référence disponible, nous comparons donc deux listes de concepts ou deux listes de termino-concepts.

Les résultats obtenus pour les deux cas d'usage présentés figurent dans la tableau 1.

Dans le cas d'AAdvantage, la référence (AA_{Ref}) est une liste de termino-concepts qui décrit le vocabulaire conceptuel des règles métier et nous avons construit deux ontologies, à partir des termes seuls (AA_1) et en prenant les entités nommées en compte dans la conceptualisation (AA_2). L'évaluation consiste donc à comparer les valeurs obtenues après l'analyse des entités nommées à l'ensemble de départ. Il existe peu d'écart entre la liste enrichie $LTCAA_2$ et celle de départ $LTCAA_1$ au niveau des valeurs précision et rappel. La précision reste stable, de 82,8% à 83%, et le rappel augmente légèrement de 67,5% à 72%. Ceci s'explique par le nombre relativement faible d'entités nommées dans le corpus. L'analyse détaillée des deux ontologies montre néanmoins que l'exploration des entités nommées dans le texte a permis de restructurer l'ontologie initiale

Résultats		Références	Mesures	
Ontologies	Liste extraites		Précision	Rappel
Ontologie AA_1	$LTCAA_1$	LTC AA_{Ref}	82,8%	67,5%
Ontologie AA_2	$LTCAA_2$		83%	72%
Ontologie $Audi$	$LCAudi$	LC $Audi_{Ref}$	72,2%	67,5%

TABLE 1: Evaluation des ontologies produites au regard de références : mesures de précision et rappel. A partir des ontologies sont extraites des listes de concepts (LC) ou de termino-concepts (LTC) qui sont comparées avec une référence.

et de l'enrichir avec de nouveaux concepts et relations. Par rapport à l'ontologie AA_1 qui contient 130 concepts, 7 nouveaux concepts ont été ajoutés dans AA_2, 15 concepts existants ont été redéfinis et 45 instances ont été ajoutées. Nous avons considéré comme bruit les entités nommées de villes car elles ne sont pas pertinentes pour l'application visée mais toutes les autres entités nommées détectées (60% du total) ont été ajoutées d'une manière ou autre à l'ontologie. Considérer les entités nommées durant la phase de conceptualisation a donc contribué à identifier des éléments pertinents du domaine (concepts) et à peupler partiellement l'ontologie.

Dans le cas d'Audi, nous avons construit une ébauche en une seule passe de conceptualisation et en nous appuyant uniquement sur les entités nommées et les termes figurant dans leur voisinage. La référence est une liste de concepts extraite de l'ontologie fournie par l'expert. Nous comparons le résultat obtenu $LCAudi$ à la référence $Audi_{Ref}$. On obtient dans ce cas d'usage une précision de 72,2% mais un rappel de 67,5%. Ces chiffres montrent l'intérêt de l'utilisation des entités nommées dans le démarrage du processus de conceptualisation : en partant des entités nommées et en explorant leurs contextes, on obtient une ébauche d'ontologie déjà proche de celle de l'expert.

Dans ces deux expérimentations relatives au domaine réglementaire, nous avons montré que l'identification des entités nommées et leur utilisation comme marqueurs durant la phase de conceptualisation guide l'ingénieur de la connaissance dans la détection et la conceptualisation d'éléments pertinents du domaine. Comme la liste d'entités nommées produite par les outils est plus réduite que celle des termes (67 entités nommées contre 634 termes pour le cas d'usage d'AAdvantage et 67 contre 511 termes pour le cas d'usage d'Audi), il est plus facile de s'appuyer sur les entités nommées pour démarrer la conceptualisation et construire une première ébauche d'ontologie. On peut ensuite itérer en explorant les contextes des unités textuelles sélectionnées précédemment et en prenant en compte les nouveaux termes figurant dans leur voisinage.

6. Conclusion

Dans cet article, nous avons montré comment une méthodologie de construction d'ontologies à partir de textes peut s'appuyer sur des unités textuelles spécifiques durant le processus d'acquisition : les termes et les entités nommées. Et nous avons expliqué le rôle particulier que les entités nommées jouent dans le processus de conceptualisation. L'approche combinée, qui est implémentée dans la nouvelle version de l'outil TERMINAE, est illustrée sur deux cas d'usage où des ontologies de domaine sont créées pour aider la formalisation des règles métier de système d'aide à la décision. Les entités nommées ne sont pas nombreuses dans les documents réglementaires, à la différence, par exemple, des articles de journaux mais nous avons montré leur importance pour l'acquisition du modèle d'un domaine donné. Même dans le cas où elles sont considérées comme des instances au niveau conceptuel, les entités nommées pointent vers des éléments caractéristiques du domaine à modéliser et jouent un rôle important dans la structuration du modèle formel.

Remerciements

Ce travail a été financé par le projet EU-IST Integrated Project 2009-231875 ONTORULE et a bénéficié de nombreuses discussions avec nos collègues F. Lévy, A. Guissé (LIPN), J. Hall (Model Systems, UK) et P. Rosina (Audi, DE).

Références

AUSSENAC-GILLES N., BOURIGAULT D., CONDAMINES A. & GROS C. (1995). How can knowledge acquisition benefit from terminology ? In *Proceedings of the 9th Knowledge Acquisition Workshop*.

AUSSENAC-GILLES N., DESPRES S. & SZULMAN S. (2008). The terminae method and platform for ontology engineering from texts. In P. BUITELAAR & P. CIMIANO, Eds., *Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text*, p. 199–223. IOS Press.

BONTCHEVA K. & CUNNINGHAM H. (2003). The semantic web : A new opportunity and challenge for human language technology. In *Proceedings of Workshop on Human Language Technology for the Semantic Web and Web Services, 2 nd International Semantic Web Conference, Sanibel Island*.

P. BUITELAAR, P. CIMIANO & B. MAGNINI, Eds. (2005). *Ontology Design and Population*. Amsterdam : IOS Press.

CIMIANO P. (2006). *Ontology Learning and Population from Text : Algorithms, Evaluation and Applications*. Springer.

CIMIANO P. & VÖLKER J. (2005). Text2onto - a framework for ontology learning and data-driven change discovery. In *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems*, p. 227–238.

FAURE D. & NÉDELLEC C. (1999). Knowledge acquisition of predicate argument structures from technical texts using machine learning : the system asium. In D. F. ET R. STUDE, Ed., *Proceedings of the 11th International Conference on Knowledge Engineering and Knowledge Management (EKAW'99)*, p. 329–334 : Springer-Verlag.

HARRIS Z., GOTTFRIED M., RYCKMAN T., PAUL MATTICK J., DALADIER A., HARRIS T. & HARRIS S. (1989). *The Form of Information in Science : Analysis of an Immunology Sublanguage*. Dordrecht : Reidel.

LERAT P. (2009). La combinatoire des termes. exemple : nectar de fruits. In *Hermes. Journal of Language and Communication Studies*, p. 211–232.

LOPES L. & VIEIRA R. (2009). Automatic extraction of composite terms for construction of ontologies : an experiment in the health care area. *Electronic Journal of Communication, Information and Innovation in Health*, **3**(1), 72–84.

MAGNINI B., PIANTA E., POPESCU O. & SPERANZA M. (2006). Ontology population from textual mentions : Task definition and benchmark. In *Proceedings of the OLP2 workshop on Ontology Population and Learning*, Sidney, Australia.

MAYNARD D., YAOYONG L. & WIM. P. (2008). Nlp techniques for term extraction and ontology population. In P. BUITELAAR & P. CIMIANO, Eds., *Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text*, p. 107–127. IOS Press.

MEYER I., SKUCE D., BOWKER L. & ECK K. (1992). Towards a new generation of terminological resources : an experiment in building a terminological knowledge base. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING'92)*, p. 956–960, Nantes, France.

NADEAU D. & SEKINE S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigaciones*, **30**(1), 3–26.

NÉDELLEC C., NAZARENKO A. & BOSSY R. (2009). Information extraction. In S. STAAB & R. STUDER, Eds., *Handbook on Ontologies in Information Systems*, chapter 31. Springer Verlag, second edition.

SEKINE S. & NOBATA C. (2004). Definition, dictionary and tagger for extended named entities. In *Proceedings of the Forth International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

TANEV H. & MAGNINI B. (2008). Weakly supervised approaches for ontology population. In *Proceeding of the 2008 conference on Ontology Learning and Population : Bridging the Gap between Text and Knowledge*, p. 129–143, Amsterdam, The Netherlands, The Netherlands : IOS Press.

WANG Y., VOLKER J. & HAASE P. (2006). Towards semi-automatic ontology building supported by large-scale knowledge acquisition. In *AAAI Fall Symposium On Semantic Web for Collaborative Knowledge Acquisition*, volume FS-06-06, p. 70–77, Arlington, VA, USA : AAAI AAAI Press.